Multi-stream Fusion for Class Incremental Learning in Pill Image Classification

Trong-Tung Nguyen[1,2], Hieu H. Pham[1,3,*], Phi Le Nguyen[4], Thanh Hung Nguyen[4], and Minh Do[1,3,5]

[1] VinUni-Illinois Smart Health Center, VinUniversity, Hanoi, Vietnam;
{tung.nt,hieu.ph,minh.do}@vinuni.edu.vn
[2] John von Neumann Institute, University of Science, VNU-HCM, Vietnam;
[3] College of Engineering & Computer Science, VinUniversity, Hanoi, Vietnam;
[4] School of Information and Communication Technology, Hanoi University of Science and Technology, Vietnam;
{lenp,hungnt}@soict.hust.edu.vn
[5] University of Illinois at Urbana-Champaign, US;minhdo@illinois.edu
*Corresponding author

Abstract. Classifying pill categories from real-world images is crucial for various smart healthcare applications. Although existing approaches in image classification might achieve a good performance on fixed pill categories, they fail to handle novel instances of pill categories that are frequently presented to the learning algorithm. To this end, a trivial solution is to train the model with novel classes. However, this may result in a phenomenon known as catastrophic forgetting, in which the system forgets what it learned in previous classes. In this paper, we address this challenge by introducing the class incremental learning (CIL) ability to traditional pill image classification systems. Specifically, we propose a novel incremental multi-stream intermediate fusion framework enabling incorporation of an additional guidance information stream that best matches the domain of the problem into various state-of-the-art CIL methods. From this framework, we consider color-specific information of pill images as a guidance stream and devise an approach, namely "*Color Guidance with Multi-stream intermediate fusion*"(CG-IMIF) for solving CIL pill image classification task. We conduct comprehensive experiments on real-world incremental pill image classification dataset, namely VAIPE-PCIL, and find that the CG-IMIF consistently outperforms several state-of-the-art methods by a large margin in different task settings. Our code, data, and trained model are available at https://github.com/vinuni-vishc/CG-IMIF.

1 Introduction

Pill image recognition task has attracted various studies recently with the aim to design high-quality algorithm for visual-based assistance system on pill images. This can help the healthcare community automatically identify unknown pill categories by taking several real-world pictures with mobile devices. It is noteworthy that real-world scenarios of pill images are often challenging due to the changing background as well as variances of pill instances in terms of shape, color, and texture. There have been several works that are developed to mitigate such challenges, most of them are based on hand-crafted features [3, 5, 6, 10]. These works are then utilized by Ling et al. [16] and combined with a two-stage training strategy to create a novel framework for the pill recognition model in few-shot learning. Another approach is to explore external knowledge from medical text data (e.g. prescription) to improve the detection performance of visual-based models [18, 19]. However, existing models are often limited by novel instances of pill categories which frequently arrive at a pill recognition system. This often happens when a novel class of pill instance is introduced by images uploaded from the end-user using mobile devices or from the healthcare community. A report in [1] shows that there are roughly 40-50 novel drugs being approved each year. In such a scenario, the core learning model of the system, which is often deployed in a lightweight device (*e.g*, mobile phones), might need to rewind the training process on the whole training data (in which novel categories participate). This is not an effective strategy for many reasons. Memory allocated for such extensively training data is often limited. Acquiring novel knowledge while maintaining what the model has learned so far requires the system to store a huge amount of samples for both old and new classes, which is infeasible. Another solution for this is to provide an initial training dataset for the model. The model is then fine-tuned on novel categories to update the model's knowledge about new pill instances. However, this fine-tuning scheme suffers from a serious behavior of the learning system which is widely known as catastrophic forgetting [8, 9] (degrading performance on old tasks while accessing data of novel tasks). This system, therefore, is in need of a flexible and effective strategy to handle the novel real-world object categorization of pill image instances. In this way, it would be able to incrementally learn from new classes without exhaustively storing old category samples. This scenario is called class-incremental learning (CIL).

The progress of studies on class incremental learning (CIL) for visual tasks has been developed significantly for many years. The general setting of CIL is that the disjoint sets of different classes arrive at the learning algorithm gradually. Many works such as [4, 13, 21–23] have proposed several methods which employed available techniques to tackle the mutual challenge: catastrophic forgetting. Knowledge distillation [12] is the most common technique which is widely adopted to tackle catastrophic forgetting and was first applied to the CIL setting by Li et al. [15]. After that, a derived version [21] with additional usage of representation learning was proposed, in which valuable herding exemplars are replayed frequently to keep track of the old knowledge. The strategy of herding is to pick those neighbors which are nearest to the mean sample of the class. Using this herding strategy, Castro et al. [4] managed to build an end-to-end framework with an additionally balanced fine-tuning strategy. On the other hand, Wu et al. [22] introduced a bias correction approach by adding a bias correction layer. This is conducted at the last layer of each incremental learning task to refine the overall scores for the final prediction. Meanwhile, Hou et al. [13] identified the imbalance between previous and new data as the main issue leading to catastrophic forgetting. They tackled this imbalanced scenario by incorporating three main components: cosine normalization, less-forget constraint, and inter-class separation.

In this research, we aim to investigate the application of CIL methods in a pill classification system. Fig.1 illustrates the effect of such a system with and without class incremental learning capability. To the best of our knowledge, we are the first to explore incremental learning on the pill classification system. Existing single stream incremental learning methods [4,13,21–23], when being applied to a domain of application for practical usage, can be improved with the help of some domain-specific knowledge. This serves as additional information which might collaborate well with the original RGB image to alleviate catastrophic forgetting. The introduction of a supplementary information stream requires a prudent strategy to incorporate such information. Based on this motivation, we propose a novel integration framework that serves as a plug-in technique for any available class incremental learning algorithms. Our fusion framework enables the incremental learning methods to receive additional information streams as cues. This will then help to flexibly update corresponding feature representations in an optimal way for each learning task through the intermediate stage. To demonstrate the usage of such an integration framework, we consider color information as additional stream and devise an approach, named "*Color Guidance with Multi-stream intermediate fusion*"(CG-IMIF). Experimental results on a real-world incremental pill image classification dataset called VAIPE-PCIL show that the proposed learning framework consistently surpasses most metric scores of various state-of-the-art methods in different task settings.

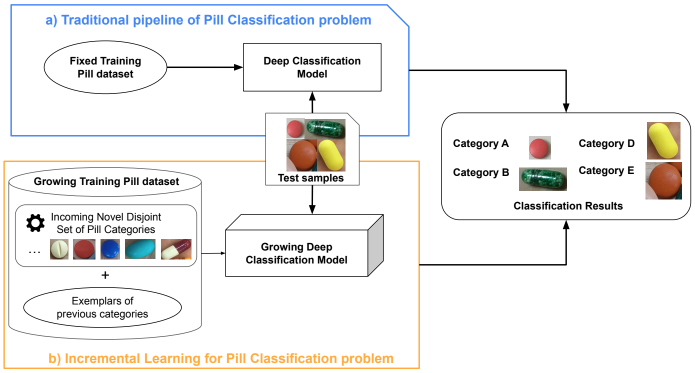

a) Traditional pipeline of Pill Classification problem
Fixed Training Pill dataset
Deep Classification Model
Test samples
Category A
Category B
Category D
Category E
Classification Results
Growing Training Pill dataset
Incoming Novel Disjoint Set of Pill Categories
...
+
Exemplars of previous categories
Growing Deep Classification Model
b) Incremental Learning for Pill Classification problem

Fig. 1: The pipeline for a learning algorithm to acquire knowledge of pill categories could be divided into two options: (a) feeding a fixed pill images database to an off-the-shelf deep learning algorithm; (b) maintaining a few samples of old categories as exemplars, combining with novel categories to form a growing pill image dataset, and finally feeding into a growing deep classification model.

Our contributions can be summarized in the following three aspects:

1. We introduce CG-IMIF, a novel incremental learning framework based on multiple streams for the task of pill classification from images. To the best of our knowledge, we are the first to introduce the incremental learning capability to this task and provide a new approach to tackle challenges in learning novel pill classes.
2. We conduct thorough experiments and in-depth ablation studies to demonstrate the effectiveness of the proposed approach on a real-world incremental pill image classification dataset. Experimental results show that the CG-IMIF consistently outperforms previous state-of-the-art methods by a large margin.

The rest of this paper is organized as follows. We briefly formulate the problem setting of pill CIL, which we aim to solve in Section 2. Details of our proposed CG-IMIF framework are described in Section 3. Experimental results and further analysis are presented in Section 4 and 5. Finally, we conclude the paper with our discussion on strengths and limitations in Section 6, and 7.

2 Preliminaries

2.1 Problem Definition and Notation

Generally, the class incremental learning (CIL) problem represented by τ consists of a sequence of n image classification learning tasks

$$\tau = [(C^1, P^1_{train}, P^1_{test}), (C^2, P^2_{train}, P^2_{test}), ..., (C^n, P^n_{train}, P^n_{test})], \tag{1}$$

where each tuple $(C^t, P^t_{train}, P^t_{test})$ depicts a task t. C^t is a set of m^t categories, *i.e.*, $C^t = \{c^t_1, c^t_2, ..., c^t_{m^t}\}$, P^t_{train} and P^t_{test} denote the training and testing data, respectively. To represent the total number of classes up to the current task, we define $M^t = \sum_{i=1}^{t} \left|C^i\right|$. The training, and testing data is defined as $P^t = \{(X^t, Y^t)\}$ where X^t and Y^t denote the training images and their corresponding labels, respectively. During the training phase, the learning model at stage t is presented with categories set C^t, training samples P^t_{train}, and an exemplar set K_t. In practice, K_t is a fixed-size set acting as a support set which helps to retain a partial set of images and the corresponding labels from previous training data, *i.e.*, $K_t \subseteq P^1_{train} \cup P^2_{train} \cup ... \cup P^{t-1}_{train}$. Therefore, a revised version of training samples at stage t can be obtained by combining K_t and P^t_{train}, $K_t \cup P^t_{train} = V^t_{train}$. It is also assumed that categories of different learning tasks do not overlap ($i.e$ $C^i \cap C^j = \varnothing$ where $i \neq j$). At testing time, the performance of learner t is evaluated on all of the previous seen categories $\bigcup_{i=1}^{t} C^t$ with samples from $\bigcup_{i=1}^{t} P^t_{test}$.

2.2 Conventional CIL Methods

Several CIL methods have been proposed which consider various properties of CIL problem to tackle mutual challenge: catastrophic forgetting. Most CIL works are divided into two branches: exemplar-based, and non-exemplar-based approaches. While the latter is much more challenging, the former is more practical since it is reasonable to maintain a few samples of old classes to avoid performance degrading. Within the scope of this research, we aim to exploit the capability of exemplar-based CIL methods by attaching these to our proposed framework. Therefore, we first describe a few core components of exemplar-based CIL approaches as follows.

Representative Memory is a set of samples from categories of old tasks and is represented by K_{t-1}. It serves as an exemplar set to support model in revisiting knowledge acquired from old tasks. In an exemplar-based approach, the learning model can only access the previous category set C^{t-1} through K_{t-1}. The size of the support set is often limited and mainly divided into two memory settings: 1) a constant number of exemplars per class, and 2) and a limited capacity of S samples. In the first setting, the size of the support set K_t grows with the number of classes. In addition, the size of K_t in the second memory setting is constant over the time t. Samples across categories are manipulated frequently with two main operations: new sample selection and old sample removal. For each class, a sorted list of its samples is maintained based on their distances to the class' mean feature vector. Hence, the most representative samples for each class are selected as members in the next support set K_{t+1}. Meanwhile, the remaining samples are ignored to reserve slots for novel samples from new classes.

Growing Deep Neural Networks in an exemplar-based method is constructed by two main factors: the common feature extractor backbone, and a growing classification layer module. At a specific learning stage t, new classification head CL_t is initiated to allocate corresponding parameters W_t. Feature vectors, after being extracted by the feature extractor, are fed into CL_t to produce prediction logits for the current category set C^t. The size of the logits after being input to CL_t is equal to the size of the category set C^t. The prediction vectors are then utilized to compute the traditional cross-entropy loss which represents the training loss on the set of pill images P^t_{train} for the current task. On the other hand, the old classification head CL_{t-1} can be used to represent the old knowledge of the model. Samples from support set K_t can be passed through a list of classification head module CL_i from the first task to the latest old task $t-1$ ($i.e$ $i \in [0, t-1]$) in order to obtain prediction logits.

Cross-Distillation Loss Function is common in most of exemplar-based methods. This is constructed by combining cross-entropy, and distillation loss function. The cross-entropy loss function helps minimize the overall empirical errors when learning on new category set C_t at task t. Meanwhile, the distillation loss function plays a role in distilling the old model M_{t-1} from previous tasks into the current model M_t to avoid catastrophic forgetting. Let's consider the incremental learning model at a specific learning stage t where it has obtained $t-1$ numbers of classification heads. New classification head CL_t, which is now added to learn on new task t, produce the prediction logits as $o(x) = [o_1(x), o_2(x), ..., o_t(x)]$ for any input x. Similarly, output logits which are produced by old classification head can be represented as $\hat{o}(x) = [\hat{o}_1(x), \hat{o}_2(x), ..., \hat{o}_{t-1}(x)]$. With these representation, the distillation loss can be computed for all samples from exemplar sets K_{t-1} and from new classes P^t_{train} (*i.e.* $K_{t-1} \cup P^t_{train} = V^t_{train}$) as follows:

$$\begin{gathered} L_d = \sum_{x \in V^t_{train}} \sum_{k=1}^{t-1} -\hat{\pi}_k(x) \log\left[\pi_k(x)\right], \\ \hat{\pi}_k(x) = \frac{e^{\hat{o}_k(x)/T}}{\sum_{j=1}^{t-1} e^{\hat{o}_j(x)/T}}, \quad \pi_k(x) = \frac{e^{o_k(x)/T}}{\sum_{j=1}^{t-1} e^{o_j(x)/T}}, \end{gathered} \tag{2}$$

where T plays as the temperature scaling factor. Meanwhile, the groundtruth label for each sample x (*i.e.* y(x)) for new category sets along with softmax of logits of the k-th category set (i.e. $p_k = softmax(o_k(x))$) can be used to compute the cross-entropy loss function as follows

$$L_c = \sum_{x \in V^t_{train}} \sum_{k=1}^{t} -y(x) \log\left[p_k(x)\right], \tag{3}$$

The final cross-distilled loss function; therefore, can be obtained by combining distillation loss function in Eqn.2 and cross-entropy loss function in Eqn.3

$$L = \alpha L_d + (1 - \alpha) L_c, \tag{4}$$

where the scalar value α controls the balance between the two functions.

3 Methodology

The majority of current pill identification methods rely on RGB images. Therefore, to the best of our knowledge, most existing systems fail to address hard examples (*e.g*, pills with very similar shapes and colors) [16]. This problem becomes more challenging in the context of the class incremental learning. In this problem, we have to cope with two issues at the same time: 1) recognizing pill instances that belong to the novel classes, and 2) not forgetting the previously learned knowledge of the old ones. We seek in this study robust domain-specific knowledge, which could be in good companion to traditional RGB image stream. However, the introduction of an additional stream issue a different challenge; the significant need for a stream integration method . To tackle such a challenge, we propose an Incremental Multi-stream Intermediate Fusion framework (IMIF). The IMIF allows additional information streams to be effectively propagated during the incremental learning phase. In the following subsections, we briefly define the multi-stream class incremental learning method and describe how it can be decomposed into different components.

3.1 Multi-stream Class Incremental Learning Model

We define a multi-stream class incremental learning model **M** as a combination of three key components: 1) a single stream base method **X**, 2) an additional stream of information **Y**, and 3) a method of fusing stream **Z**.

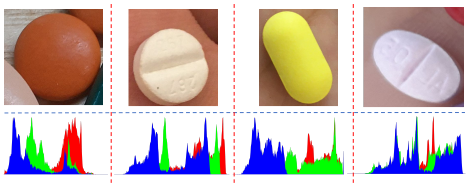


Fig. 2: Samples of pill images from VAIPE-PCIL dataset are shown on the first row. The second row exhibits the corresponding color histogram information for images on the upper row respectively.

$$\mathbf{M} = \text{Base method } \mathbf{X} + \text{Feature stream } \mathbf{Y} + \text{Fusion mechanism } \mathbf{Z}$$

At this point, the base method represents any method that follows the general setting described in Sec.2.2. **Y** serves as a piece of additional domain information that gives cues to the learning model apart from RGB images. Normally, **Y** is specific to the domain of the task. Lastly, **Z** presents a fusion mechanism that enables method **M** to incorporate additional information stream **Y** into the incremental learning process. From this decomposition, our CG-IMIF replaces: 1) the representative stream **Y** with color-specific information, and 2) the fusion technique **Z** with the proposed IMIF. In the following, we describe our proposed Color histogram guidance stream and Incremental Multi-stream Intermediate Fusion technique in Sec.3.2 and Sec.3.3, respectively.

3.2 Color Histogram Guidance Stream

Pill images compose of various features which can be used as discriminative factors in classification problems. Among those is color distribution information which can be approximated by the color histogram of pill images. The color histogram represents the color distribution of the input image in terms of three different channels: red, green, and blue. In detail, it is often encoded into a single vector where the indexes of entries are mapped to a set of all possible colors. In grey-scale images, values of vector entries store the frequency that counts the total number of pixels having the intensity (*i.e.*, color value), which is hashed by the corresponding index of the vector. However, in a three-dimensional image, color ranges for each channel are associated across color channels to formulate a unique combination. This combination accounts for those pixels of which color ranges lie in three discrete ranges of value: $[r_i, r_j]; [g_k, g_l]; [b_m, b_n]$. The color feature vectors are useful to make a distinction among pill instances that have similar shapes but different colors. The color ranges for each channel of the RGB images are divided into eight segments in our problem, where each segment represents 32 different consecutive color values. After that, the color histogram vector can be obtained by accumulating the quantities of pixels assigned to a specific color range for each channel to result in a vector with $8 \times 8 \times 8 = 512$ elements. Fig. 2 illustrates a few examples of extracting the color histogram stream for each corresponding cropped pill image.

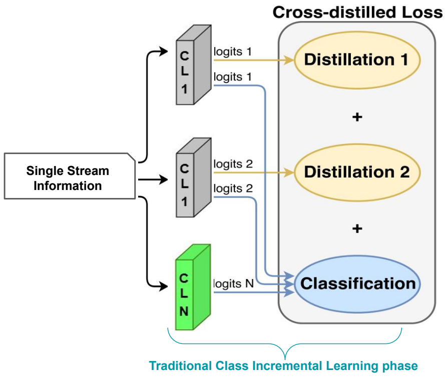


Fig. 3: The traditional training paradigm with only single-stream information used by almost existing exemplar-based CIL methods.

3.3 Incremental Multi-stream Intermediate Fusion Technique

Traditional Early Fusion A naive fusion technique $\mathbf{Z}$ concatenates different streams of information right after the feature extraction phase. Specifically, feature vectors $f_r \in \mathbb{R}^{d_r}$, and $f_Y \in \mathbb{R}^{d_Y}$ are extracted from raw RGB images, and additional information stream $\mathbf{Y}$, respectively. Both of these features are then fed into separate projection layers to project into the same latent space. In practice, the projection layers are implemented by a single hidden layer controlled by parameters $\Theta^p = [W^p, b^p]$ as follows

$$s_r = \sigma(W_r^p.f_r + b_r^p), s_Y = \sigma(W_Y^p.f_Y + b_Y^p), s_g = [s_r, s_Y]. \tag{5}$$

The projected vectors are then concatenated to obtain the global feature vector $s_g \in \mathbb{R}^{d_g}$. This global feature s_g can be considered as a single input into any traditional single stream CIL methods as shown in Fig. 3.

Intermediate Fusion We observed that fusing information stream in an early manner for a class incremental learning problem is not optimal. The global feature s_g is problematic since it is regularly updated at each incremental task. As a result, the projection layer in the early phase can not find good parameters that balance the performance of old and novel categories in different tasks. Therefore, we propose to relocate the projection layer to the intermediate phase instead ($i.e$, incremental learning phase) by initiating an entirely novel projection layer in an incremental manner. When C^t from new task t arrives at the system, a new classification layer CL_t is created . This layer accompanies by an attached projection layer specific to the information stream of a specific task. Therefore, the parameters controlling the projection layer for each information stream are different from those defined in the early fusion.

$$\Theta_r^{p^t} = [W_r^{p^t}, b_r^{p^t}], \Theta_Y^{p^t} = [W_Y^{p^t}, b_Y^{p^t}]. \tag{6}$$

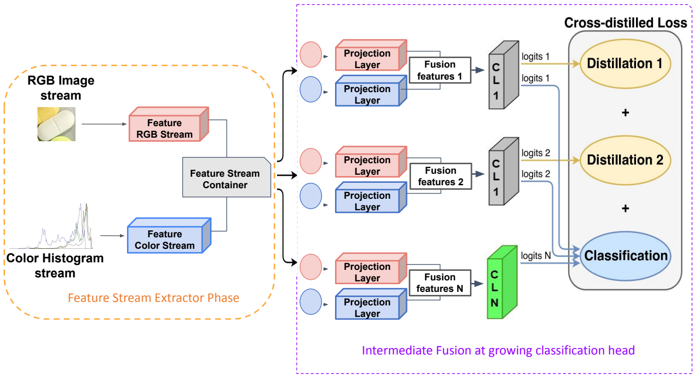


Fig. 4: Our proposed CG-IMIF architecture composes of: 1) color histogram feature extraction (orange block), and 2) intermediate fusion framework (purple block) to incorporate additional information stream.

4 Experiments

4.1 Dataset

We employ a real-world image dataset, namely VAIPE-Pill (VAIPE Pill Identification) [2] to exploit CIL capability on the pill image classification problem. This dataset is created to promote the research on recognizing distinct types of medicines from mobile devices. The dataset contains 7,294 pill images of 262 categories taken in real-world scenarios. The characteristics of VAIPE-Pill dataset are illustrated in Tab. 1.

To facilitate research of CIL in pill image classification tasks, we derive a dataset version, namely VAIPE-PCIL (VAIPE Pill Class Incremental Learning) dataset from the original VAIPE-Pill data. VAIPE-PCIL is obtained by cropping pill instances from the original data. We only select those categories which satisfy either of the following conditions: 1) the number of samples should not be too small (*i.e.*, and 2) larger than 10 samples), image size of samples should be at least 64 × 64. Samples of pill image from VAIPE-PCIL can be found in Fig. 1. All of our experiments are conducted on the VAIPE-PCIL dataset to study the performance of CG-IMIF.

4.2 Experimental Protocol

Settings We follow the standard benchmark protocol proposed in [13]. We fix class arrangements in random order. After each training stage t, the resulting learner is evaluated on the testing data $\bigcup_{i=1}^{t} P_{test}^{t}$ which represents for all of the testing data up to the current task t. Since no test data from the previous learning stage are hidden from the learner, it is guaranteed that no overfitting can occur.

Table 1: Statistics of VAIPE-Pill dataset on different characteristics.

Characteristic	Training set	Testing set	Total
Number of images	6,461	833	7,294
Number of pill categories	262	262	262
Instances per category	179.75	23.56	203.2
Image size (pixel×pixel, mean)	$3,311 \times 3,276$	$3,276 \times 3,469$	$3,300 \times 3,400$
Instances per image	7.28	7.4	7.3
Number of bounding box annotations	47,097	6,174	53,271
Number of categories per image	5.18	5.76	5.32

There are two commonly different task evaluation settings in class incremental learning: task-awareness and task-agnostic. The first setting is much easier for the algorithm since it has access to the task ID (*i.e.*, ID or set of categories) about the incoming test data. Therefore, it is reasonable to only use the corresponding classification head in the incremental learning phase, which is trained on that task-ID to evaluate the performance. This task setting, however, is not practical in many real-world circumstances since task-ID is not always available. We evaluate our performance in terms of task-agnostic instead. In task-agnostic, the model is not given the task identities of the test data. Hence, the evaluation results are achieved by taking the results of all prediction logits, which are predicted by all of the classification heads. In this way, the model has to learn to resolve the confusion among classes from a different set of classes.

Evaluation Metrics We adopt two commonly used benchmark metrics from [13] for CIL problems: average accuracy and average forgetting rate. The average accuracy and forgetting rate records of performance for each incremental learning phase are often if a single number is preferable. Meanwhile, the average phase accuracy and forgetting rate would be used to observe learning behaviors during incremental tasks for each method.

4.3 Implementation Details

All of our experiments and methods are implemented with Pytorch [20] and trained on a single NVIDIA GeForce RTX 3090. We inherit the codebase from FACIL [17]. They have already implemented various state-of-the-art methods for CIL problems in a well-structured manner. Details of base models as well as the implementation of our IMIF framework are discussed below. In all experiments, we attach our IMIF framework to several state-of-the-art methods in CIL: BiC [22], EEIL [4], and LUCIR [13]. Since these methods followed the common prototype of exemplar-based methods, we discussed some of the general settings of exemplar-based methods in our experiments before diving into details about the setting of each base one. There are two common strategies to reserve samples for old classes: 1) exemplar-management stores a constant number of samples for each old class, or 2) it maintains a fixed capacity (*e.g*, $R_{total} = 2,000$ for CIFAR-100 [14] and $R_{total} = 20,000$ for ImageNet [7]). In our experiments, we follow the first setting since it is usually more challenged than the second one.

In addition, the exemplars are randomly selected among different categories. For the training network, we made use of 50-layer ResNet [11] with no pre-trained weights as the backbone for feature extraction module, which is applicable to the VAIPE-PCIL dataset. To fairly compare the performance, we fixed the number of training epochs (200 epochs) across different methods. The learning rate is initialized with 0.1 and is divided by 1.5 if the loss function suffers from non-decreasing circumstances for a specific number of attempts ($e.g$, $lr_{patience} = 5$). The networks are trained using stochastic gradient descent with mini-batches of 32 samples. The training images are resized to the same shape of $256 \times 256 \times 3$ with only one transformation ($e.g$, flipping). The class orders across different methods are randomly fixed for a fair comparison.

In terms of configuration for base methods, we follow the same settings for the original version BiC [22], and our improved version BiC-CG-IMIF. BiC [22] proposed to integrate a bias correction layer attached to the end of each classification head to adjust the classification score. The number of training epochs for the bias correction layer is 200 epochs in our setting. Moreover, we set 0.1 as the ratio of the number of exemplars that are used for the validation. EEIL [4] performs an additional fine-tuning phase after each official training phase to balance the performance between old and novel categories. In our experiments, we fix 40 as the number of epochs for fine-tuning and the learning rate fine-tuning factor as 0.01 across different methods. We also adhere to the base setting of LUCIR [13] method where they removed ReLU in the penultimate layer to take both positive and negative values. For the IMIF framework, the projection layer implemented is represented by a single hidden layer. Therefore, the output size for different projection layers should be the same so that the transformed feature vectors, then can be fused in the shared space. In terms of the color-guided information, color ranges for each channel of the RGB images are divided into 8 segments where each segment represents 32 different consecutive pixel values.

4.4 Experimental Results

We evaluate our proposed CG-IMIF approach and report the overall performance in comparison with several state-of-the-art approaches in Tab.2 Experimental results show that most of the state-of-the-art approaches attached with our proposed IMIF tool and color-specific information as additional stream help to achieve consistent improvements over task settings. The setting consists of three tasks in total where the number of categories is uniformly distributed for 5, 10, and 15 tasks. It is noticeable that the lower score of forgetting rate indicates that the model is more unlikely to forget about old knowledge. In addition, average phase accuracy and forgetting rate is also illustrated in Fig. 5,6 to inspect the learning behaviour of each method through incremental phases. Dashed and solid lines with different colors are utilized to differentiate the base ones (X) and our CG-IMIF, respectively. In terms of the average phase accuracy, LUCIR-CG-IMIF obtains the highest performance where it can consistently and significantly surpass other methods (also the base one-LUCIR). On the other hand, BiC-CG-IMIF is better at mitigating the forgetting constraint. However,

Table 2: Average accuracy $\bar{\mathcal{A}}$ (%) and forgetting rate $\mathcal{F}$ (%) of CG-IMIF compared to other state-of-the-art results in different task settings. Best scores are marked in bold for both evaluation metrics.

Metric	Method	*Task Settings*		
		N=5	N=10	N=15
Average acc. **(%)** ↑ $\bar{\mathcal{A}} = \frac{1}{n}\sum_{i=1}^{n}\mathcal{A}_i$	EEIL [4]	63.83	62.40	57.41
	EEIL-CG-IMIF	**70.80**	**64.85**	**60.93**
	BiC [22]	53.83	55.75	53.77
	BiC-CG-IMIF	**65.53**	**63.59**	**54.83**
	LUCIR [13]	69.63	62.90	55.49
	LUCIR-CG-IMIF	**76.85**	**69.94**	**64.97**
Forgetting rate. **(%)** ↓ $\bar{\mathcal{F}} = \frac{1}{n}\sum_{i=1}^{n}\mathcal{F}_i$	EEIL [4]	49.82	45.46	48.27
	EEIL-CG-IMIF	**46.68**	**44.64**	**46.23**
	BiC [22]	20.05	30.50	26.93
	BiC-CG-IMIF	**7.75**	**22.01**	**27.35**
	LUCIR [13]	44.13	44.32	47.11
	LUCIR-CG-IMIF	**33.15**	**37.88**	**39.79**

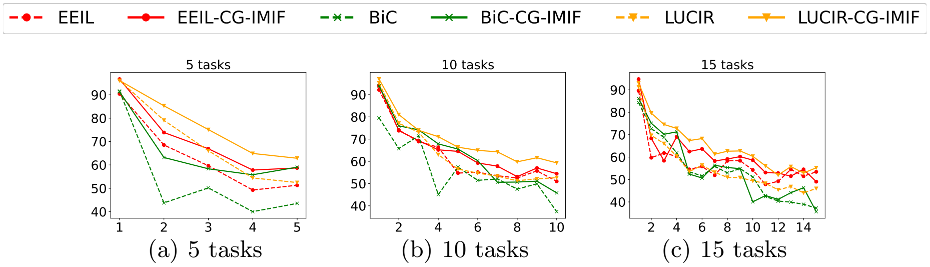
$^{\diamond}$ Using the similar exemplar settings and selection for fair comparison.

it is not consistent over tasks and the curve fluctuates. One possible explanation is that the bias layer inside the traditional BiC method and BiC-CG-IMIF might cause the model to sacrifice the performance of the current task to maintain the memory of the old ones.

5 Ablation Studies

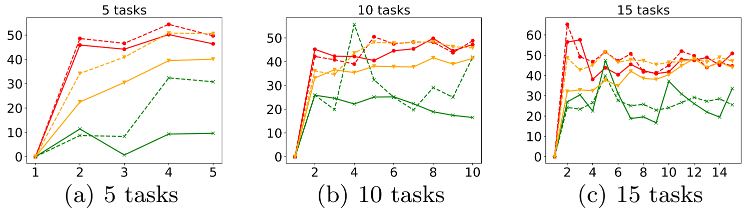
To examine the effect of additional information stream usage and fusion framework, we perform extensive ablation studies. This is aim to observe the effect of different components in our proposed framework where LUCIR is chosen as the base method. LUCIR is preferable because of the consistency and high performance of LUCIR across task settings in the experimental results which have been discussed in Sec. 4.4. In addition to color information, edge signals might be a good candidate to discriminate different pill categories based on their shape. To understand the importance of different stream usage in our method, we compare 4 different settings: 1) RGB image only, 2) RGB and edge images, 3) RGB and color histogram, and 4) a combination of all three streams. Each separated row in Tab.3 refers to each scenario of information stream usage with two different fusion techniques. Concretely, the setting that combines RGB and color histogram streams achieves the highest score. One possible explanation for this result is that the edge signal might not be sufficiently strong to push the performance.

In addition, we implement the basic fusion technique where additional information streams are fused in an early manner. Each separated row in Tab.3



(a) 5 tasks (b) 10 tasks (c) 15 tasks

Fig. 5: Incremental accuracy for different task settings among the original version and our method CG-IMIF.



(a) 5 tasks (b) 10 tasks (c) 15 tasks

Fig. 6: Incremental forgetting rate for different task settings among the original version and our method CG-IMIF.

illustrates the results of two different fusion mechanisms. Our fusion technique (on the second line of each row) outperforms the traditional one in various metrics and task settings. The best result is LUCIR-CG-IMIF which integrates color histogram information into the traditional LUCIR method with IMIF.

6 Discussions

Key Findings. To the best of our knowledge, this work is the first to tackle the class incremental learning problem for the pill image domain, which is crucial and applicable for real-world pill recognition systems. Also, we empirically showed that the technique of intermediate fusion with the additional stream is superior to the early fusion technique. One plausible explanation for this effect is the flexibility of the fusion layer after it has been relocated to the intermediate stage. This allows the additional information to maintain its optimal performance for old tasks while learning to adapt to new tasks.

Limitations. Though the proposed framework has superior performance over several state-of-the-art methods in CIL, it contains some limitations in different aspects. The new fusion layer at the intermediate phase might enlarge the model's size in terms of the number of parameters. Considering the scenario when a massive amount of tasks are encountered in the learning progress, the learning model could create sequences of abundant layers. This might create a side effect

Table 3: Ablation performance to compare variants of combination which utilize different information streams as well as different fusion techniques. The combination which achieves highest performance over different tasks is our CG-IMIF and is marked in green.

Variant of Combination	*Average acc.* (%) ↑			*Forgetting rate.* (%) ↓		
	N=5	N=10	N=15	N=5	N=10	N=15
RGB only	69.93	62.90	55.49	44.13	44.32	47.1
RGB-Edge + Early	70.94	63.90	55.28	42.4	42.13	45.80
RGB-Edge + Intermediate	**72.58**	**68.38**	**62.90**	**38.78**	**38.19**	**41.02**
RGB-Color + Early	73.58	64.57	53.56	37.825	42.86	46.15
RGB-Color+ Intermediate	76.85	69.94	64.97	33.15	37.88	39.79
RGB-Edge-Color + Early	69.99	63.33	56.34	42.35	44.17	46.24
RGB-Edge-Color+ Intermediate	**73.65**	**68.32**	**62.15**	**36.30**	**38.48**	**40.58**

when too much memory is reserved for storing the model's parameters. Such reservation is unreasonable in a real-world deployment. Another restriction with the proposed framework is related to additional stream utilization. Apart from the traditional RGB stream, another information channel that is specific to the domain of usage might impose a disagreement with the original RGB channel. This requires a careful study of a different combination of streams accompanying the traditional stream to observe its effect.

7 Conclusion

This paper introduces the incremental learning capability to the traditional pill image classification systems. To this end, we propose a novel framework, namely Incremental Multi-stream Intermediate Fusion (IMIF) which integrates an additional stream of information to improve the performance of the single stream CIL method. We then devise CG-IMIF which utilizes IMIF along with a color histogram as guidance information. Our CG-IMIF is flexible and can be attached to any exemplar-based approach to improve the performance of the base ones. We experimentally show that CG-IMIF outperforms many existing state-of-the-art methods on the VAIPE-PCIL dataset. We hope our work would lay the foundation and could benefit several types of future research into the continual learning ability of intelligent machines in smart health applications.

Acknowledgements. This work was funded by Vingroup Joint Stock Company (Vingroup JSC),Vingroup, and supported by Vingroup Innovation Foundation (VINIF) under project code VINIF.2021.DA00128. Trong-Tung Nguyen was funded and and supported by the Master, PhD Scholarship Programme/Post-Doctoral Scholarship Programme of Vingroup Innovation Foundation (VINIF), Vingroup Big Data Institute (VinBigdata), code VINIF.2021.ThS.JVN.07.

References

1. CDER's New Molecular Entities and New Therapeutic Biological Products — fda.gov. https://www.fda.gov/drugs/development-approval-process-drugs/new-drugs-fda-cders-new-molecular-entities-and-new-therapeutic-biological-products, [Accessed 04-Jul-2022]
2. Anh Duy, N., Dung Thuy, N., Thanh Hung, N., Phi Le, N., Hieu H., P., Minh N., D.: VAIPE: A Large-scale and Real-World Open Pill Image Dataset for Visual-based Medicine Inspection. https://vaipe.org/
3. Caban, J.J., Rosebrock, A., Yoo, T.S.: Automatic identification of prescription drugs using shape distribution models. In: 2012 19th IEEE International Conference on Image Processing. pp. 1005–1008 (2012). https://doi.org/10.1109/ICIP.2012.6467032
4. Castro, F.M., Marín-Jiménez, M.J., Guil, N., Schmid, C., Alahari, K.: End-to-end incremental learning. CoRR **abs/1807.09536** (2018), http://arxiv.org/abs/1807.09536
5. Chen, Z., Kamata, S.i.: A new accurate pill recognition system using imprint information **9067** (11 2013). https://doi.org/10.1117/12.2051168
6. Chen, Z., Yu, J., Kamata, S.i., Yang, J.: Accurate system for automatic pill recognition using imprint information. IET Image Processing **9** (07 2015). https://doi.org/10.1049/iet-ipr.2014.1007
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
8. French, R.: Catastrophic forgetting in connectionist networks. Trends in cognitive sciences **3**, 128–135 (05 1999). https://doi.org/10.1016/S1364-6613(99)01294-2
9. Goodfellow, I.J., Mirza, M., Xiao, D., Courville, A., Bengio, Y.: An empirical investigation of catastrophic forgetting in gradient-based neural networks (2013). https://doi.org/10.48550/ARXIV.1312.6211, https://arxiv.org/abs/1312.6211
10. Hartl, A.: Computer-vision based pharmaceutical pill recognition on mobile phones (05 2012)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR **abs/1512.03385** (2015), http://arxiv.org/abs/1512.03385
12. Hinton, G.E., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. CoRR **abs/1503.02531** (2015), http://arxiv.org/abs/1503.02531
13. Hou, S., Pan, X., Loy, C.C., Wang, Z., Lin, D.: Learning a unified classifier incrementally via rebalancing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
14. Krizhevsky, A., Nair, V., Hinton, G.: Cifar-100 (canadian institute for advanced research) http://www.cs.toronto.edu/~kriz/cifar.html
15. Li, Z., Hoiem, D.: Learning without forgetting. CoRR **abs/1606.09282** (2016), http://arxiv.org/abs/1606.09282
16. Ling, S., Pastor, A., Li, J., Che, Z., Wang, J., Kim, J., Le Callet, P.: Few-shot pill recognition. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9786–9795 (2020). https://doi.org/10.1109/CVPR42600.2020.00981
17. Masana, M., Liu, X., Twardowski, B., Menta, M., Bagdanov, A.D., van de Weijer, J.: Class-incremental learning: survey and performance evaluation. arXiv preprint arXiv:2010.15277 (2020)

18. Nguyen, A.D., Nguyen, T.D., Pham, H.H., Nguyen, T.H., Nguyen, P.L.: Image-based contextual pill recognition with medical knowledge graph assistance. arXiv preprint arXiv:2208.02432 (2022)
19. Nguyen, T.T., Nguyen, H.D., Nguyen, T.H., Pham, H.H., Ide, I., Nguyen, P.L.: A novel approach for pill-prescription matching with gnn assistance and contrastive learning. arXiv preprint arXiv:2209.01152 (2022)
20. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc. (2019), http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
21. Rebuffi, S., Kolesnikov, A., Lampert, C.H.: icarl: Incremental classifier and representation learning. CoRR **abs/1611.07725** (2016), http://arxiv.org/abs/1611.07725
22. Wu, Y., Chen, Y., Wang, L., Ye, Y., Liu, Z., Guo, Y., Fu, Y.: Large scale incremental learning. CoRR **abs/1905.13260** (2019), http://arxiv.org/abs/1905.13260
23. Zhang, J., Zhang, J., Ghosh, S., Li, D., Tasci, S., Heck, L.P., Zhang, H., Kuo, C.J.: Class-incremental learning via deep model consolidation. CoRR **abs/1903.07864** (2019), http://arxiv.org/abs/1903.07864